Committee on Data Science

The Committee on Data Science (CDS) was established in 2023 to support graduate and undergraduate programs in this emerging discipline at the University of Chicago. Affiliated faculty come from numerous departments across campus with a core group in the departments of Statistics and Computer Science. CDS holds the educational philosophy that a strong program in Data Science should encompass foundational theory, methodological innovations and real-world applications. A Data Science education should draw from the intellectual tradition and key concepts of Computer Science, Applied Mathematics, Statistics, and other fields while providing a new integrative framework for data-driven thinking, discovery, and decision-making.

Committee on Data Science website: https://codas.uchicago.edu/

Committee Co-Directors

- Dan L Nicolae (Statistics)
- Michael J Franklin (Computer Science)

Program Faculty

- Luc Anselin (Sociology)
- Luis Bettencourt (Ecology, Sociology)
- Raul Castro Fernandez (Computer Science)
- Aloni Cohen (Computer Science)
- James Evans (Sociology)
- Nick Feamster (Computer Science)
- Robert Grossman (Medicine, Computer Science)
- Ari Holtzman (Computer Science, Data Science)
- Nikos Ignatiadis (Statistics, Data Science)
- Alex Kale (Computer Science, Data Science)
- Frederick Koehler (Statistics, Data Science)
- Sanjay Krishnan (Computer Science)
- Mina Lee (Computer Science, Data Science)
- Bo Li (Computer Science, Data Science)
- Tian Li (Computer Science, Data Science)
- Sendhil Mullainathan (Computation, Behavioral Science)
- Samantha Riesenfeld (Molecular Engineering, Medicine)
- Veronika Rockova (Econometrics, Statistics)
- Aaron Schein (Statistics, Data Science)
- Matthew Stephens (Statistics)
- Chenhao Tan (Computer Science, Data Science)
- David Uminsky (Computer Science)
- Blase Ur (Computer Science)
- Victor Veitch (Statistics, Data Science)
- Jingshu Wang (Statistics)
- Molly Offer-Westort (Political Science)
- Rebecca Willett (Statistics, CAMI, Computer Science)
- Haifeng Xu (Computer Science, Data Science)
- Ce Zhang (Computer Science, Data Science)

PhD in Data Science

Program Overview

The PhD in Data Science was developed to train all students in the mathematical foundations of data science, responsible data use and communication, as well as advanced computational methods. Candidates will be able to explore diverse research opportunities alongside distinguished Data Science faculty at UChicago.

Curriculum

The program requires students to complete nine courses: four required courses (1-4 below); one elective either in mathematical foundations or scalability and computing (5 or 6 below), and four other graduate-level electives that can come from proposed courses in Data Science or existing graduate courses in Computer Science.
or Statistics. Some students, after consulting with the committee graduate advisor, might decide to take all nine courses over the first two years.

Required Courses:
1. Foundations in Machine Learning & AI - Part I
2. Responsible Use of Data & Algorithms
3. Data Interaction
4. Systems for Data and Computers / Data Design

Required Electives (Choose one of the following):
1. Foundations in Machine Learning & AI - Part II
2. Data Engineering & Scalable Computing

Thesis Advisor and Dissertation Committee

Students typically select a thesis advisor by the beginning of their second year. By the end of the third year, each PhD student shall establish a thesis committee of at least three faculty members, including the advisor, with at least half of the members coming from the Committee on Data Science (CDS).

Proposal Presentation & Admission to Candidacy

By the end of the third year, students should have scheduled and completed a proposal presentation to their committee in order to be advanced to candidacy. The proposal presentation is typically an hour-long meeting that begins with a 30-minute presentation by the student followed by a question and discussion period with the committee.

Admissions

The PhD in Data Science admits students each year for the Fall quarter only; a full list of admission requirements and a link to start your application can be found here (https://codas.uchicago.edu/how-to-apply/). If you have any questions regarding your application or the admissions process, please send your inquiry to data-science@uchicago.edu for a timely response.

M A S T E R ' S I N D ATA S CIENCE (MSDS)

Program Overview

The Master’s in Data Science (MSDS) was developed for students interested in pursuing a research career in Data Science with courses taught by faculty in Statistics, Computer Science, and other departments across the university.

Curriculum: Foundational Courses

The program offers three foundational courses. Students have the option to either (1) enroll in foundational courses in the summer before the program starts or (2) pass examinations to demonstrate proficiency in the material in lieu of enrolling in foundational courses.

The foundational courses are as follows:
1. Computational Foundations for Data Science
2. Mathematical Foundations for Data Science
3. Statistical Foundations for Data Science

Curriculum: Core & Elective Courses

In addition to the foundational courses (or passing examinations in lieu of enrollment in foundational courses), students must complete five required core courses, four graduate-level electives (approved by the Committee on Data Science), as well as a final project in order to be eligible for degree completion.

The core courses are as follows:
1. Introduction to Data Science
2. Systems for Data and Computers/Data Design
3. Data Interaction
4. Introduction to ML and AI or Foundations of Machine Learning and AI - Part I
5. Responsible Use of Data and Algorithms

Admissions
DATA 30332. Thinking with Deep Learning for Complex Social & Cultural Data Analysis. 100 Units.
A deluge of digital content is generated daily by web-based platforms and sensors that capture digital traces of human communication and connection, and complex states of society, culture, economy, and the world. Emerging deep learning methods enable the integration of these complex data into unified social and cultural "spaces" that enable new answers to classic social and cultural questions, and also pose novel questions. From the perspective of deep learning, everything can be viewed as data-novels, field notes, photographs, lists of transactions, networks of interaction, theories, epistemic styles-and our treatment examines how to configure deep learning architectures and multi-modal data pipelines to improve the capacity of representations, the accuracy of complex predictions, and the relevance of insights to substantial social and cultural questions. This class is for anyone wishing to analyse textual, network, image or arbitrary structured and unstructured data, especially in concert with one another to solve complex social and cultural analysis problems (e.g., characterize a culture; predict next year's ideology).
Instructor(s): James Evans Terms Offered: Spring
Prerequisite(s): The course uses Python and the widely popular PyData ecosystem to demonstrate all motivating examples and includes working code, accompanying exercises, relevant datasets and additional analytics and visualization that facilitate social and cultural interpretation and communication. Familiarity with Python is required.
Equivalent Course(s): MACS 27000, MACS 37000, SOCI 30332

DATA 33221. Advanced Topics in Law and Computing. 100 Units.
This interdisciplinary seminar will bring together instructors and graduate students from Computer Science / Data Sciences and the Law School. The seminar's focus will be on topics where law and policy intersect with computer science. Such topics may include cryptography and encryption; electronic surveillance and criminal procedure; the Computer Fraud & Abuse Act; the law governing data breaches; redistricting and the US Census; deep fakes; GDPR, Europe's Digital Services Act and the CCPA; and international data transfers. Students will be evaluated on the basis of short bi-weekly reaction papers, class participation based on weekly assigned reading, and team projects that pair law students with computer and data scientists.
Equivalent Course(s): CMSC 33221

DATA 35422. Machine Learning for Computer Systems. 100 Units.
This course will cover topics at the intersection of machine learning and systems, with a focus on applications of machine learning to computer systems. Topics covered will include applications of machine learning models to security, performance analysis, and prediction problems in systems; data preparation, feature selection, and feature extraction; design, development, and evaluation of machine learning models and pipelines; fairness, interpretability, and explainability of machine learning models; and testing and debugging of machine learning models. The topic of machine learning for computer systems is broad. Given the expertise of the instructor, many of the examples this term will focus on applications to computer networking. Yet, many of these principles apply broadly, across computer systems. You can and should think of this course as a practical hands-on introduction to machine learning models and concepts that will allow you to apply these models in practice. We'll focus on examples from networking, but you will walk away from the course with a good understanding of how to apply machine learning models to real-world datasets, how to use machine learning to help computer systems operate better, and the practical challenges with deploying machine learning models in practice."
Instructor(s): Nick Feamster
Prerequisite(s): CMSC 14300 or CMSC 15400
Equivalent Course(s): DATA 25422, CMSC 25422, CMSC 35422

DATA 37711. Machine Learning 1. 100 Units.
This course is a graduate level introduction to machine learning. We will cover both practical and (probabilistic) foundational aspects of the subject. Topics include empirical risk minimization and overfitting, regression, ensembling, as well as selected topics on representation learning and structure learning.
Instructor(s): V. Veitch Terms Offered: Autumn
Prerequisite(s): Open to graduate students in data science, statistics, and computer science, or by instructor consent. Students should have a strong background in programming, basic probability, and a graduate-student level of mathematical maturity.
Equivalent Course(s): CAAM 37711, STAT 37711

DATA 37784. Representation Learning in Machine Learning. 100 Units.
This course is a seminar on representation learning in machine learning. The core questions in this are: how do machine learning systems recover the structure present in real-world data, how can we expose this recovered structure to human analysts, and how does this help us in real-world applications? In this seminar, we will read and discuss papers from the modern research literature on these subjects. Students should have previous exposure to machine learning and deep learning.
Terms Offered: TBD
Equivalent Course(s): STAT 37784